This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

Assessing the Representation of Suicidal Ideation in Social Media Datasets Relative to Suicide Notes

Merve Kavut, Amina Dzafic, Ulya Bayram*

***Abstract*—Clinical records are inaccessible due to legal restrictions, and genuine suicide notes are scarce, limiting the availability of verified suicidal expression data. Consequently, most Natural Language Processing (NLP) studies rely on widely accessible social media datasets containing expert annotations or heuristic labels. However, it remains unclear how expressions of suicidal ideation in these datasets relate to the ideation in genuine suicide notes. In this study, we address this gap by comparing manually annotated social media datasets in English and Turkish, as well as automatically labeled English datasets, directly against genuine suicide notes using a combination of linguistic and statistical analyses, classification models, and zero-shot embedding-based representations. Results show that expert-labeled social media data exhibit limited overlap with suicide notes across both languages while remaining distinguishable, whereas automatically labeled data diverge substantially from the notes in all analyses. These findings indicate that the annotation method critically shapes the signals learned by the computational models and that social media data capture different stages or forms of ideation than those in suicide notes. Overall, our findings emphasize the importance of careful dataset evaluation and caution against misinterpreting model performance in suicide risk detection research, given the shared objective: supporting global suicide prevention efforts.**

***Index Terms*—NLP, suicide notes, social media, affective computing, text classification, zero-shot embeddings, suicidal ideation detection.**

I. Introduction

Suicide remains a leading cause of death worldwide, posing a critical public health challenge that demands scalable and proactive intervention efforts [1], [2]. As populations increasingly turn to digital platforms to express their thoughts and emotions, researchers have been focusing on developing NLP frameworks to detect suicidal ideation signals. These developments promise to help mental health professionals reach larger populations, providing early detection and intervention, especially in under-resourced regions where clinical support is limited. Consequently, automated suicidal ideation detection has become a key research area. However, the success of the NLP models in achieving this goal depends on the availability and quality of reliable training data. The gold standard clinical records are largely inaccessible due to privacy, legal, and ethical restrictions [3]–[5]. Similarly, genuine suicide notes, which offer invaluable insights into the mindset at the final stage of ideation [6], [7], can be collected from public sources [8], but are insufficient for the state-of-the-art Large Language Models (LLMs), which typically require large volumes of data [9], [10]. Consequently, researchers rely on any accessible data source, which is mostly social media [3]–[5], [11].

Among the public social media datasets, a limited portion are expert-annotated for suicidal ideation levels [12], [13]. However, it remains unclear how expressions of suicidal ideation in these datasets relate to the ideation in genuine suicide notes, which were produced at the final stage of ideation. The lack of this comparison raises fundamental questions: Do expert annotations on social media data contain signals of suicidal ideation similar to those in the notes, or do they reflect qualitatively different stages or forms of distress? Finding the answers to these questions would impact how we interpret suicide risk detection models trained on these social media datasets. Since most datasets and studies are in English, with only a few studies conducted in Spanish, Korean, or Chinese [14]–[17], models and findings derived from English data may not generalize to global populations with different linguistic and cultural backgrounds. In that case, how much experimental results from expert-annotated datasets align across languages remains an open question. Further, a more significant uncertainty arises when considering the case of automatically labeled social media datasets, which are typically constructed using heuristic criteria such as keyword matching or their domain origin [18], [19], e.g., labeling all posts collected from the r/suicidewatch subreddit as suicidal. Despite their widespread use, the relationship between these auto-assigned labels and verified suicidal expressions, i.e., the notes, remains unexplored, raising questions: Are these auto-assigned labels reliable? Should they be used in suicide-related NLP research?

Addressing these questions is vital for refining the interpretation of NLP model outputs trained on social media data, considering their intended real-world applicability in suicide prevention. In this study, we address them through a range of statistical, computational, and representational analyses, and provide several novel contributions to the domain of suicide ideation detection:

1. **Statistical Benchmarking of Social Media Datasets:** There is a prevalent reliance on Linguistic Inquiry, and Word Count (LIWC) features for interpreting suicidal ideation patterns in text [20]–[22]. Influenced by this reliance, a study used LIWC features to compute Cohen's d effect size to decide whether to merge genuine suicide notes with another dataset [23]. In this study, we replicate this methodology to benchmark the notes against an expert-labeled social media corpus, an auto-labeled social media corpus, and a corpus from an unverified source. Consequently, we measure the linguistic similarity between the notes and these corpora and examine

All authors are with Department of Electrical and Electronics Engineering, Çanakkale Onsekiz Mart University, Çanakkale, Türkiye. *Corresponding author

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

how different annotation strategies affect the findings.

2. **Computational Benchmarking:** The literature has shifted towards using machine learning approaches [12], [24], recently followed by LLMs [5], [25]. In our framework, we conduct the same comparative assessments across notes and social media data using traditional machine learning, deep learning models, and LLMs. Their results allow us to assess whether the computationally determined alignment or divergence is consistent with the previous statistical benchmarking outcomes.

3. **Multilingual Evaluation in English and Turkish:** To address the question of whether patterns observed in English social media data and suicide notes generalize to other languages, we extend our framework to Turkish, a language spoken by over 80 million people. We collect a novel Turkish suicide notes dataset and a manually annotated Turkish social media dataset, and repeat the computational analyses we applied in English, including n-gram comparisons, classification experiments, and zero-shot embedding visualizations. This multilingual evaluation allows us to assess whether the alignment and divergence patterns observed in English are also present in Turkish, and to identify language-specific differences in the representation of suicidal ideation.

The remainder of this paper is organized as follows. Section II summarizes related work, including data collection practices and concerns regarding dataset reliability in suicide research. Section III describes the selected datasets and outlines the methods. Section IV presents the experimental results. Finally, Section V discusses the findings, interprets their implications for suicide risk detection, and outlines directions for future.

II. Related Work

In this section, we review studies that aim to detect suicidal ideation from textual data, containing genuine suicide notes or content collected from social media platforms. We intentionally exclude studies conducted on clinical data as they are outside the scope due to the difficulty of accessing such data. Table I summarizes datasets and the methods employed in these studies together with their performance scores.

One of the early studies conducted using genuine suicide notes collected 33 genuine suicide notes and 33 simulated notes to compute their distinguishability, and used machine learning methods including Logistic Regression (LR) and Decision Trees (DT), achieving 78% accuracy [6]. Then, mental health professionals and psychiatric trainees performed the same task, which resulted in 63% and 49% accuracy, respectively. Through these results, they highlighted the power of computational models in suicide research. The first large-scale genuine suicide notes dataset consisted of 1,319 notes collected between 1950 and 2011 by Dr. Edwin Shneidman [7].

Influenced by the previous efforts, a study constructed the Genuine Suicide Notes (GSN) corpus, collecting 142 notes written between 1958 and 2016 from online resources such as news articles, books, and diaries in English [8]. These notes were classified against love letters and texts from a depression project using simple classifiers, achieving an accuracy of 78% [8]. The GSN corpus was later extended to 211 notes with the same data collection strategy and then classified against depressive social media posts from Reddit, and neutral blog entries using a modified Long Short-Term Memory (LSTM), achieving an F1 score of 82.7% [26].

Building on the GSN corpus, a study incorporated a dataset called Kaggle Suicide Notes (KSN), creating a larger combined collection of 659 notes [23]. Before performing this combination, they compared the GSN and KSN datasets using LIWC 2015 features and Cohen's d effect size method and reported no significant difference between their suicidal expressions. Subsequently, they classified the GSN+KSN dataset against 431 last statements from Texas death-row inmates (1982-2017) and 2,000 neutral Reddit posts using a Bidirectional LSTM (BiLSTM)-based transformer, reporting F1 scores above 90%. An alternative study created the CEASE corpus, which contained 2,393 sentences from 205 suicide notes, labeled by emotions [31]. Later expanded to 2,539 sentences by adding 120 new notes, forming CEASEv2.0 [32].

Unlike the small-scale suicide notes collections, social media platforms offer significantly more data. Further, they contain dynamic and contemporary thoughts and expressions that could help to capture early signs of distress. Thus, many recent studies relied on social media data for developing models to detect suicidal ideation. A study scraped 8 million posts from 15 mental health subreddits, including r/SuicideWatch, r/schizophrenia, and r/depression, among others [12]. Next, they selected 500 users from this collection at random, four practicing psychiatrists assessed these users based on their posts following the Columbia-Suicide Severity Rating Scale (CSSRS), and assigned one of five predetermined labels to each user. Next, through experimenting with term-frequency inverse-document-frequency (tf-idf) features and machine learning methods, including Support Vector Machines (SVM), Random Forest (RF), and deep learning methods such as Feed-Forward Neural Networks (FFNN) and Convolutional Neural Networks (CNN), they achieved an overall 65% F1 score in detecting the CSSRS labels of the users. This collection is publicly available, thus used by many other studies. One of them reported 79% F1 on this corpus using Bidirectional Encoder Representations from Transformers (BERT) models, Multilayer Perceptrons (MLP), and Suicide Ideation Severity Modeling with Ordinal loss (SISMO) [27]. Another study reported a 72% F1 score using LSTM, BiLSTM, and MentalBERT on this corpus while also classifying another expert-labeled Reddit dataset, and proposed Robust Early-exiting Internal Classifiers for Suicide Risk Evaluation (RISE) and Selective Abstention for Suicidality Identification (SASI) methods, among others [28].

There are many studies in the literature that used heuristically labeled datasets. One of the popular ones is the Suicide and Depression Detection (SDD) corpus containing posts from r/SuicideWatch, r/depression, and r/teenagers subreddits [18]. A study excluded the depression posts and used SDD for binary classification of suicidal and non-suicidal posts based on their subreddit origin [19]. Using Universal Sentence Encoders (USE) with CNNs, while experimenting with LSTMs and machine learning methods including eXtreme Gradient Boosting (XGBoost or XGB), they reported 94.2% F1 score.

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

TABLE I
SUMMARY OF THE SELECTED RELATED WORK THAT USES SUICIDE NOTES AND/OR SOCIAL MEDIA DATA IN ENGLISH FOR SUICIDE RESEARCH.

Study	Data	Methods	Score
[6]	33 genuine, 33 simulated notes	LR, DT	78% accuracy
[8]	GSN, love letters, depression notes	WEKA toolkit with LMT, J48, NB, Zero-R	86.61% accuracy, 82.66% F1
[26]	GSN, depression subreddit, neutral posts	LSTM	96.1% F1
[23]	GSN, KSN, last statements, neutral Reddit posts	BiLSTM-based transformer	above 90% F1
[12]	CSSRS	SVM (linear, rbf), RF, FFNN, CNN	71% precision, 60% recall, 65% F1
[27]	CSSRS	SVM (linear, rbf), RF, MLP, Contextual CNN, SDM, ContextBERT, SISMO, a hybrid method (Longformer + TensorGCN + Transformer Encoder + Ordinal Classification)	81% precision, 76% recall, 79% F1
[28]	CSSRS, CLPsych2022	RISE, SASI, MentalBERT, SISMO, n-BiLSTM, LSTM, ContextBERT, SDM, Contextual CNN	CSSRS 72%, CLPsych 59% F1
[19]	SDD	USE + FCNN, SVM, XGB, LSTM, LSTM-CNN	94.16% F1
[29]	SDD, CEASEv2.0, SWMH	BiLSTM, BERT, FastText, GloVe, Sentence Embeddings, SVM, RF, DT, LR, CNN, LSTM, GRU, RCNN, XGB	SDD 97% F1, CEASEv2.0 75% F1, SWMH 68% F1
[25]	SDD	RoBERTa-CNN, RoBERTa, CNN-BiLSTM, XGB, EFS + pBGSK + SVM	98.0% accuracy, 96.8% F1, 97.6% AUC
[4]	SDD, CEASEv2.0, tweet sentiment	CNN, BERT, BERT + CNN, LSTM, Att-BiLSTM+CNN, SVM	97% F1, 97% accuracy
[30]	SDD	n-grams, GBT, NB, SVM, MentalBERT, textCNN, RoBERTa, MEDLINE/PubMed BERT	98% F1
[5]	SDD	(Word2Vec, GloVe, FastText) + C-BiLSTM, GPT2, BERT	97.69% F1
[3]	SDD, tweets	BERT, MUSE, LLM, PBFS, BiGRU + attention, CNN, XGB	Twitter 97%, SDD 98% AUC

The SuicideWatch and Mental Health collection (SWMH) is another dataset, containing posts from r/SuicideWatch, r/depression, r/Bipolar, r/Offmychest, among others, and is available from Hugging Face [33], labeled by source subreddit. From classifying the labels in the SDD and SWMH corpora, the study achieved F1 scores of 97% and 68%, respectively, using methods such as BERT, Gated Recurrent Units (GRU), and pre-trained embeddings [29]. Many studies that classified the SDD dataset in Table I reported around 97% F1 scores from a range of LLMs [3]–[5], [25], [30].

III. MATERIALS AND METHODS

A. Data

Following the previous studies [8], [26], we collected 419 genuine suicide notes in English from public resources, including news articles, books, and academic publications. Subsequently, we selected three publicly available English social media datasets. The first one is the aforementioned CSSRS dataset, comprising Reddit posts from r/SuicideWatch, manually labeled by mental health professionals [12]. It contains five labels: users who have thoughts of suicide and preoccupations with risk factors are assigned the Suicidal Ideation (SI) label, those with a higher suicide risk with a confession of active or historical self-harm or planning have the Suicidal Behavior (SB) label, Actual Attempt (AA) labels were given to those with intentional actions that could result in death, Suicide Indicator label for the users with non-suicidal supportive posts that use at-risk words, and Supportive label for the remaining users. We selected users with ideation that consist of users with SI (171 users, 2,746 posts), SB (77 users, 1,789 posts), and AA (45 users, 752 posts) labels, yielding 5,287 posts. We excluded the remaining data as they are outside the current scope. The second dataset is the heuristically labeled SDD collection [18], [19]. We randomly sampled 5,500 out of 116,037 suicidal posts from SDD to match the scale with the CSSRS dataset. The third dataset, KSN[1] is selected due to its aforementioned similarity with notes [23].

For the Turkish-language experiments, we collected 288 genuine suicide notes from public resources. Next, for benchmarking, we used a dataset with 7,874 posts we previously scraped from Eksi Sozluk[2] and annotated by a novel framework [Redacted for blind review]. Two human annotators initially assigned two major and four minor labels ({Positive, Mixed} $\in$ suicidal, {Negative, Other} $\in$ non-suicidal) to each post. Positive label represents the indisputable presence

[1]https://www.kaggle.com/datasets/mohanedmashaly/suicide-notes/
[2]https://eksisozluk.com/

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

of suicidal thoughts, Mixed represents suicidal ideation was present in addition to statements against suicide, Negative shows opposition to suicide and promoting help, and Other indicates unrelated remarks. Two annotators disagreed on 17% of the major labels, which concerns the presence of suicidal ideation; so the third human/expert annotator's labels are assigned to these posts. To resolve minor label disagreements, we obtained two LLMs (ChatGPT-4.o and Gemini 2.5) label predictions, and used the majority vote between all labels as the final decision. Only 4.8% of the posts lacked a majority vote, which then received randomly assigned labels as a form of regularization. We refer to the collection with Positive labels as TRP, the Mixed set as TRM, and their combination as TRPM. We exclude the posts with Negative or Other labels as they are outside the scope.

TABLE II
DATASETS, THEIR LANGUAGES, THE NUMBERS OF SUICIDAL TEXTS, AND THEIR AVERAGE LENGTHS AS THE NUMBER OF WORDS.

Dataset	Language	# of texts	Avg. length
Notes EN	English	419	93.0
CSSRS	English	5,287	73.4
SDD	English	5,500	198.8
KSN	English	467	157.4
Notes TR	Turkish	288	83.3
TRP	Turkish	2,986	55.17
TRM	Turkish	2,431	84.59
TRPM	Turkish	5,417	68.4

B. Statistical Method

Following a study [23], we used LIWC features and Cohen's d scores for statistical benchmarking experiments. LIWC is a psycholinguistic tool used to determine emotions or psychological states in a text by associating words with a set of categories, widely used in mental health research [22], [34]. Cohen's d is a standard measure of effect size, computed as the difference between two means divided by a weighted average of the standard deviations of the groups [35]. It has two components: magnitude and direction. High magnitudes indicate linguistic divergence, while values close to zero imply similarity between the two datasets. The sign indicates which group has the larger mean. In this study, we utilized the public LIWC 2007 [36], which has 63 feature categories in English, limiting our statistical experiments to this language. When computing the d scores, we used the notes as the first element; therefore, positive scores indicate that notes have higher mean values compared to the other datasets.

C. Computational Methods

As shown in Fig. 1, we designed our study as a binary 5-fold cross-validation classification framework, where genuine suicide notes are labeled 1, and social media posts with a suicidal label are 0. Through fixed random seeds, the splits were consistent across experiments. All models were implemented in Python 3.11.13, using Scikit-learn (version 1.6.1) for machine learning [37], PyTorch (version 2.6.0+cu124) for deep learning and transformers [38], and Optuna (version 4.2.1) for hyperparameter tuning [39]. We performed these experiments across Google Colab (machine learning), Kaggle GPUs (parts of deep learning), and a local server (remaining deep learning and transformer experiments), after ensuring compatibility and reproducibility with Python library versions.

1) Machine Learning Models: We experimented with a range of traditional classifiers with n-gram ($n \leq 2$) features. At each fold in Fig. 1, we applied ANOVA for feature selection, tested the top $k \in 250, 500, 1000, 2500, 5000$ features with LR to select the best feature subset returning the highest macro F1 on cross-validating the training set, and used this feature subset for training the machine learning models and evaluate the test set. Because the SDD and KSN corpora do not contain user identifiers, we could not cluster posts by authors; so we treated each post as an independent instance. Yet, in CSSRS and Turkish corpora, user identifiers were present. In TRP, 2,236 users contributed only a single post, while 293 users produced more than one post; among these, over 200 users authored exactly two posts. Similarly, in TRM, out of 2,118 users, 211 posted more than once, and more than 160 of these users had only two posts. Thus, we only pursued user-based training-test splitting to avoid data leakage in the CSSRS corpus in our experiments. Each machine learning model was trained with balanced class weighting or optimized sample weights to address class imbalance (see Table II).

Naive Bayes (NB): To address class imbalance, we used a Gaussian Naive Bayes classifier and performed a search over the smoothing parameter $\alpha \in [10^{-3}, 1.0]$.

LR: We searched for optimal solvers (lbfgs, saga, liblinear) and $C \in [10^{-4}, 10^{2}]$.

SVM: For linear SVM (LSVM), we searched for the optimal regularization $C \in [10^{-4}, 10^{2}]$; for the non-linear variants, we searched for $kernel \in \{rbf, polynomial, sigmoid\}$, $C \in [10^{-3}, 10^{2}]$, $\gamma \in [10^{-4}, 10]$ for rbf or sigmoid, $degree \in [2, 5]$ for polynomial.

GB: Comprises a series of weak learners trained in succession. We searched the number of estimators $\in [50, 1000]$ and learning rate $\in [0.01, 0.3]$.

RF: We searched for the number of estimators $\in [50, 1000]$, and maximum tree depth $\in [3, 30]$.

XGBoost: It has a complex decision-making mechanism, where each tree corrects the errors of the previous one by optimizing a loss function [40]. We searched for the number of estimators $\in [50, 1000]$, learning rate $\in [0.01, 0.3]$, maximum tree depth $\in [3, 10]$, minimum child weight $\in [1, 10]$, $\gamma \in [0, 5]$, subsample ratio of columns when constructing each tree $\in [0.5, 1]$, scale positive weight $\in [0.5, 5]$, and regularization parameters $\alpha \in [0, 10]$, $\lambda \in [0, 10]$.

2) Deep Learning Models: Among the popular deep learning models in the NLP literature, we selected GRU and bidirectional GRU (BiGRU), given their strength in sequential tasks. In both models, we first initialized their embedding layers with the averaged weights of the final four layers of the selected pre-trained BERT models for the relevant languages (see the Transformer section), followed by two GRU or BiGRU layers, dropout, and a classification head with two neurons. For GRU layers, we searched for the number of

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

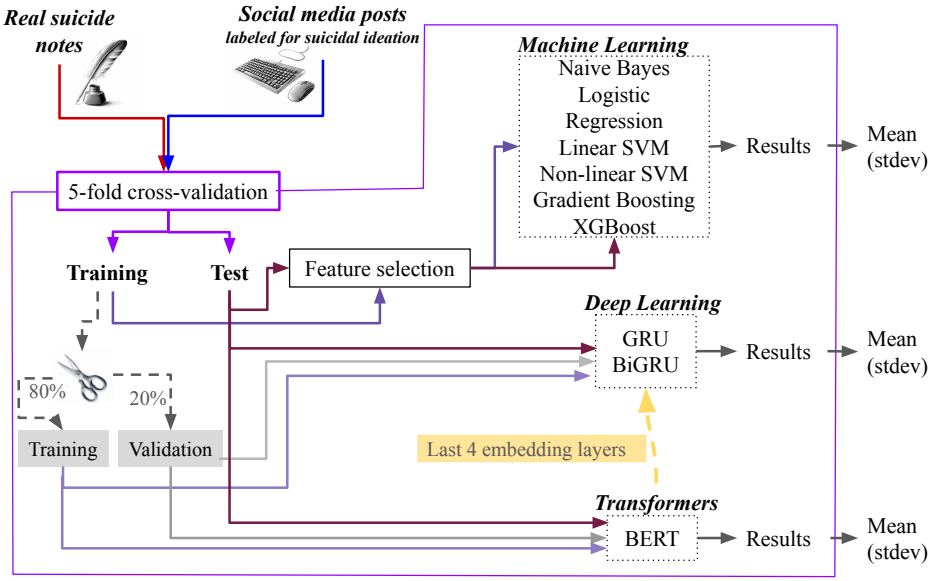


Fig. 1. Computational 5-fold cross-validation framework used to evaluate suicidal ideation detection in social media posts.

neurons $\in [64, 512]$, dropout with the rate $\in [0.1, 0.5]$, learning rate $\in [10^{-5}, 10^{-2}]$, patience for early stopping $\in [3, 10]$, and weight decay $\in [0, 0.1]$, and used the AdamW optimizer with the focal loss function ($\gamma = 2$ and $\alpha =$ division of training set class weights). BiGRU used the identical parameter search, except for a different learning rate ($\in [10^{-5}, 0.005]$) and a patience ($\in [2, 10]$). We separated validation sets for these parameter searches, partitioned from the 20% of the training data at random, and not used elsewhere. The CSSRS experiments had the validation data separation at the user-level.

3) Transformer Models: We fine-tuned two pre-trained transformer models. We selected the popular BERT-base-uncased for English experiments [9] and BERTurk for Turkish data [10], based on its prior performance on Turkish text classification [41]. To each BERT, we added dropout and a classification head. We used maximum sequence lengths derived from the average text length in the training set with a padding margin of 10 additional tokens. Fine-tuning included early stopping based on validation loss per quarter-epoch checkpoint using the same validation sets in the deep learning experiments. We searched for the optimal learning rate $\in [10^{-6}, 10^{-4}]$, weight decay $\in [10^{-6}, 10^{-2}]$, dropout rate $\in [0.1, 0.5]$, and patience $\in [3, 10]$. We used the AdamW optimizer and employed the same focal loss function as the deep learning experiments, but with $\alpha = 1$, as validation experiments showed that BERTs inherently managed class imbalance.

We also employed the transformers to evaluate their baseline ability to represent the data semantically in a zero-shot setting to avoid training bias. Therefore, we passed each document (i.e., post or note) through the selected pre-trained transformer and extracted the [CLS] token embedding from the final hidden layer. This embedding serves as a fixed-length representation of the entire input document. Then, we projected these high-dimensional embedding vectors into two dimensions using Uniform Manifold Approximation and Projection (UMAP). It is a nonlinear dimensionality reduction technique that preserves both local and global structure of the input, making it well-suited for visualizing complex relationships in embedding spaces [42]. UMAP allowed us to explore how the map onto the zero-shot embedding space.

IV. Results

A. LIWC Analyses

Using LIWC features and Cohen's d method, we compared English suicide notes with three public social media datasets. Comparing the notes with the expert-annotated CSSRS dataset yielded an average d score of 0.102, indicating a negligible linguistic difference between the two. The positive sign indicates that the notes have higher LIWC feature values than the CSSRS dataset. In contrast, comparing the SDD and KSN against notes produced larger average effect sizes, -0.288 and -0.270, respectively, where the negative values mean these datasets had higher LIWC feature values than the notes.

In Fig. 2, individual LIWC categories further present notable patterns. Categories related to cognitive processes (cogmech), including discrepancy (discrep) and causation (cause), as well as negative emotions (negemo), "time," "anger," "health," sadness (sad), and anxiety (anx), all follow the same trend: notes are linguistically similar to CSSRS with only negligible, positive differences, implying a slightly higher presence of features from these categories in the notes. Yet, the pattern is reversed for the SDD and KSN experiments, where they returned larger negative effect sizes, indicating substantially greater presence of these features in SDD and KSN relative to the notes. The first-person singular pronoun category, "i," also had the same trend, where the comparison between notes and CSSRS yielded a positive effect size, whereas SDD and KSN experiments returned negative effect sizes. This pattern aligns

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

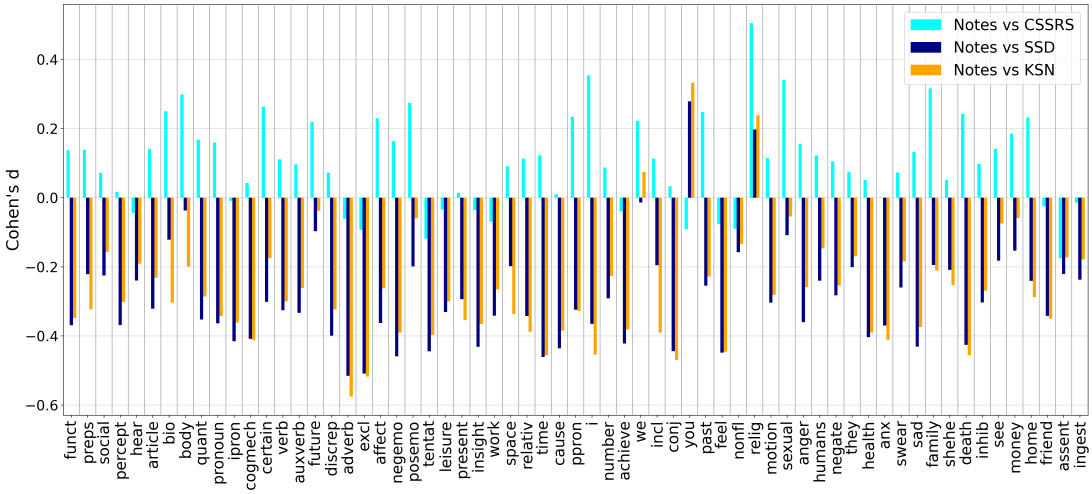


Fig. 2. Cohen's d scores obtained for the LIWC 2007 features, comparing English suicide notes with three datasets (CSSRS, SDD and KSN).

with prior clinical findings that reported significant differences in pronoun usage between suicidal and non-suicidal individuals [43]–[45]. However, this pattern breaks in the second-person pronoun category, where CSSRS had negligibly higher feature values than notes, but the SDD and KSN experiments had the opposite case.

There are other notable trends in Fig. 2. For example, in the "religion" category, all effect sizes are positive. The CSSRS versus notes comparison showed a large effect size, while the values were smaller in the SDD and KSN experiments. This observation indicates consistently higher usage of religious terms in notes compared to these datasets. Likewise, the "feel" category had negative effect sizes across all experiments, indicating more expressions of feelings in all these datasets than in the notes. Because the free version of LIWC 2007 does not support Turkish, we were unable to repeat the evaluations in Turkish-language.

B. N-gram Analyses

To characterize the feature-level variability before evaluating the machine learning results in distinguishing notes from the social media data, we conducted n-gram overlap analyses using UpSet plots across folds and experiments [46]. These analyses are helpful to achieve a more interpretable assessment of our downstream machine learning results.

The first UpSet analysis examined n-gram overlaps across cross-validation folds. To visually inspect some overlapping features, we included the n-grams above the bars containing fewer than six features in Fig. 3. The plots in this figure show that the number of n-grams shared across all folds in all experiments ranges from 36 to 61, indicating a degree of feature stability. When we analyze individual plots, the CSSRS versus notes plot in Fig. 3a contains many n-grams containing "you," in addition to "heaven, grieve, hurt, exhausted, love, goodbye, please, pray" and "my life." SDD versus notes' UpSet plot also contains "you" in many n-grams and "god forgive, darling, grieve, mercy, please, gratitude" as other notable features in Fig. 3b. The KSN versus notes plot also includes "you," and other noteworthy n-grams such as "forgive me, depressed, my parents, parents, alive, feels like" in Fig. 3c.

UpSet plots also revealed notable findings from the Turkish-language experiments. For example, Fig. 3f shows that TRPM versus notes experiments contain overlapping n-grams across folds that translate as "farewell, want to be buried, with free-will, wife, debt." Meanwhile, TRP experiment's plot emphasizes financial burden and family-related content with "my kids, bury, wife, cannot pay, got debt, my son, debt" in Fig. 3d, and TRM experiment has "forgive me, my son, wife, free-will, debt, family, cannot pay, got debt, buried" in Fig. 3e. These patterns that were absent in the English-language plots suggest that an older demographic is present in the Turkish datasets than those in the English experiments, possibly married, suicidal in the context of economic hardship, and family-oriented thoughts. Yet, a theme observed across two languages is asking for forgiveness, suggesting a universal emotional dimension of suicidal ideation, despite socioeconomic and sociodemographic differences.

The second UpSet analysis in Fig. 4 examines n-gram overlap across experiments to determine whether the same lexical features were responsible for distinguishing suicide notes from the social media corpora. Given that suicide notes in each language-specific experiment were constant, this analysis becomes crucial because, in each language, experiments use the same genuine suicide notes dataset. For each plot, we aggregated the n-gram features that survived the feature selection process across all cross-validation folds, obtaining a complete set of features for each experiment. Through these UpSet plots, we quantified the number of features that were in both experiments, and those that were unique to each.

In Fig. 4a, the CSSRS and SDD experiments share 3,358 features, corresponding to 27% of their merged feature set, with SDD exhibiting slightly more unique features than CSSRS. Similarly, in Fig. 4c, the SDD and KSN experiments share 3,430 features, which also constitute 27% of their merged feature set, and SDD again shows a slightly larger number of unique features. In contrast, comparison of the

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

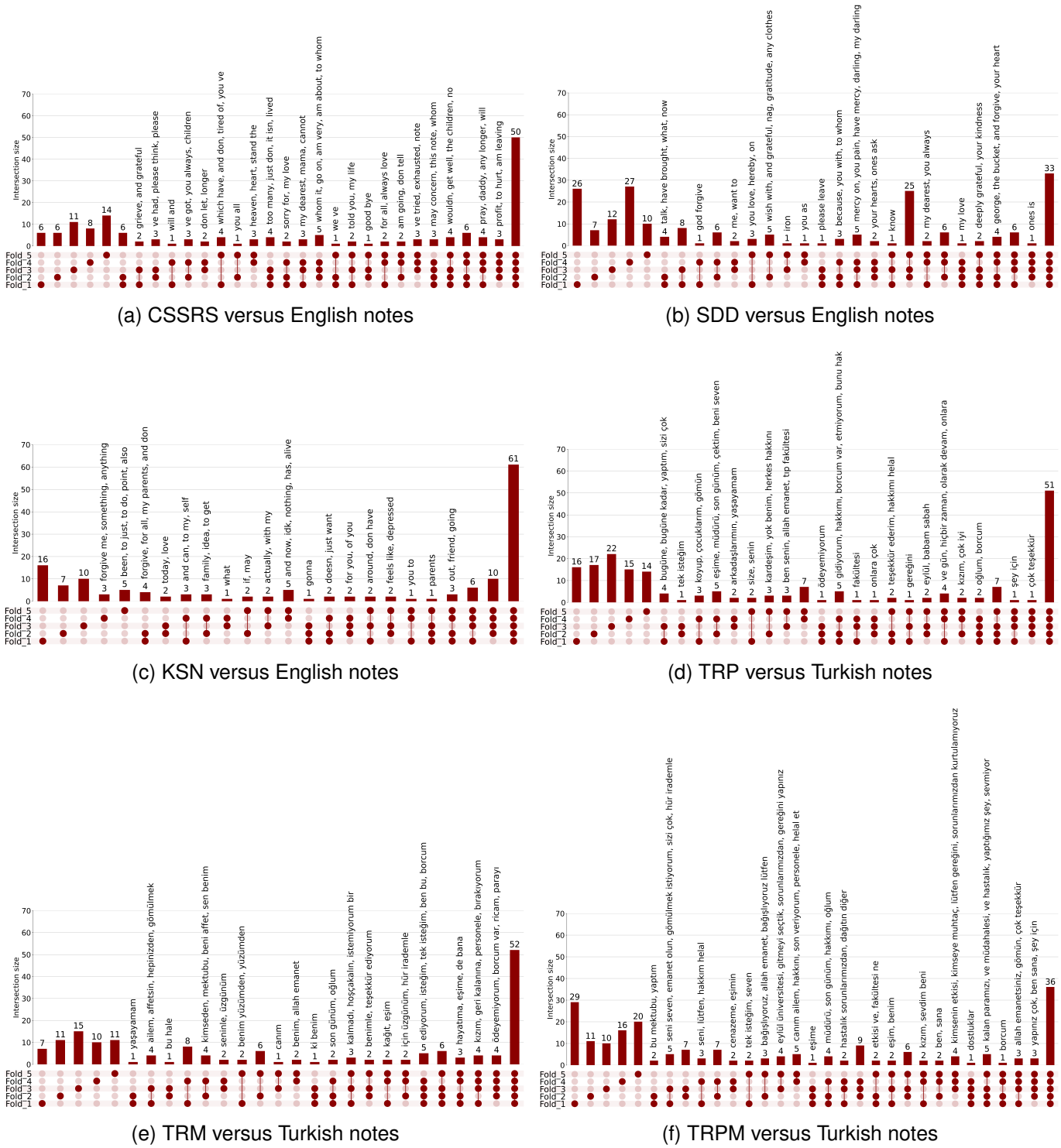


Fig. 3. UpSet plots of English and Turkish experiments, demonstrating the n-gram feature overlaps across five-fold cross-validation.

CSSRS and KSN experiments in Fig. 4b reveals a smaller shared set of 2,692 features, with the KSN experiment using a larger number of unique features. These results show that while some n-grams are consistent across experiments in English, discriminative features are dataset-specific, indicating that models rely on different lexical signals across social media corpora despite the constant suicide note data.

The Turkish social media datasets had manual annotations for two suicidal ideation levels; their UpSet analyses reveal variation in linguistic patterns when the data source and the annotators are held constant, and only the dataset labels differ. Under that perspective, the TRP and TRM experiments displayed greater n-gram overlap than their combination, TRPM, in Fig. 4d. This observation indicates internal feature consis-

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

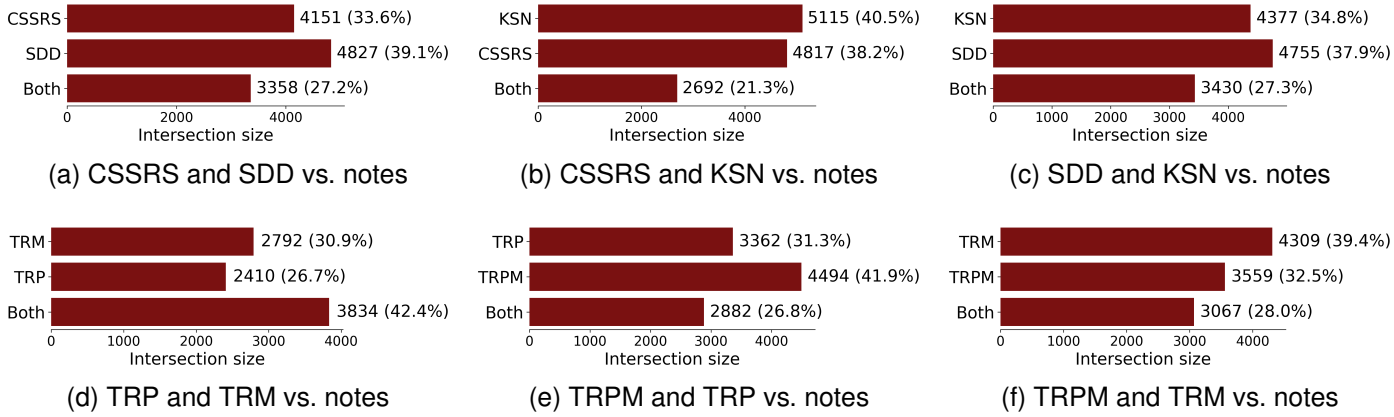


Fig. 4. UpSet plots comparing n-gram overlaps across datasets.

tency between each label-specific corpus, but the decrease in the number of overlapping features in the TRPM corpus defies intuitive expectations.

C. Classification Results

To assess the separability of suicide notes from the selected datasets, we used a wide range of computational methods. Given the class imbalance between the social media corpora and the notes in Table II, we selected macro-averaged F1, AUC, precision, and recall metrics to ensure fair evaluation across classes. Table III and Table IV reveal a consistent pattern of near-perfect AUC scores followed by lower F1 scores. According to the literature [47], AUC can overestimate model performance in imbalanced classification scenarios due to its threshold-independent ranking formulation. Therefore, we focused on the macro-averaged F1 scores as informative metrics for interpreting our classification results.

Table III shows that BERT consistently achieved the highest F1 scores across the English-language experiments. In the expert-annotated CSSRS versus notes classification task, the highest F1 score is 82.4%. Meanwhile, traditional machine learning results showed conservative predictions, characterized by higher precision and lower recall, whereas the deep learning models and the transformer returned higher recall scores at the expense of precision. We observed the same pattern in some of the SDD experiments' results, with a maximum F1 score of 83.9%. In contrast, the KSN experiments showed more balanced precision-recall trade-offs across different models, with the best-performing model achieving an F1 score of 86.0%.

Subsequently, the results of the Turkish-language experiments in Table IV show that the highest achieved F1 score was 80%, from the TRM scenario by both the linear machine learning model, LSVM, and the BERT model. The TRP dataset had an F1 score of 76.2% while the TRPM dataset had a similar performance of 76.5%, both with LSVM. Across the Turkish-language experiments, performance differences between linear models and BERT were small, showing that the classification problem was linearly separable.

To further analyze the classification results, we compared the predictions with expert annotations using confusion matrices. In the CSSRS experiment, BERT achieved the highest scores in distinguishing social media posts from suicide notes, and LSVM in the TRPM versus notes experiment, so the matrices contain the predictions of these two models. Because the primary interest in this analysis was to understand how the expert-annotated ideation levels align with model predictions, Fig. 5a and Fig. 5b show only the social media data, excluding the notes.

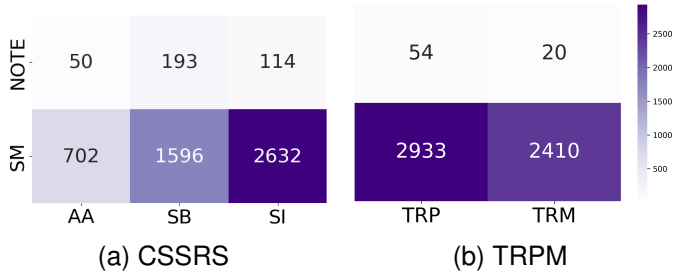


Fig. 5. Confusion matrices of the CSSRS and TRPM experiments' results against their original expert-annotated labels.

Fig. 5a shows that the majority of the posts predicted as suicide notes were from the users with the SB label. Given the higher prevalence of the SI-labeled users compared to the SB ($171 > 77$) and their number of posts ($2,746 > 1,789$) in the CSSRS dataset, this finding aligns with the definition of SB as the ideation level with a high suicide risk. In contrast, posts of the users with the AA label had the least number of predictions as notes, further validating the distinction between the ideation levels describing past suicide attempts and those reflecting the final ideation. In Fig. 5b, repeating the same analysis for the Turkish-language experiments showed that most of the social media posts were classified as social media posts. Meanwhile, among the small number of posts predicted as notes, TRP posts were predicted as notes more than twice as often as TRM posts, further validating the finding from the English-language experiment.

Next, to leverage the user-level information within the CSSRS dataset, we performed another post-analysis of its BERT predictions. For each user, we obtained the ratio of

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

TABLE III
MEAN AND STANDARD DEVIATION OF THE 5-FOLD CROSS-VALIDATION RESULTS FOR THE ENGLISH-LANGUAGE EXPERIMENTS.

Model	Precision	Recall	F1	AUC
NB	**91.0 (1.7)**	78.6 (1.8)	**83.5 (1.7)**	94.8 (1.9)
LR	80.5 (4.7)	83.2 (3.5)	81.5 (3.6)	94.4 (2.2)
LSVM	80.6 (4.6)	83.1 (3.4)	81.6 (3.4)	94.2 (2.4)
SVM	78.6 (6.0)	79.9 (2.0)	78.9 (4.2)	91.9 (1.7)
GB	**87.0 (2.9)**	70.9 (1.0)	76.3 (1.3)	91.8 (2.0)
RF	66.4 (6.8)	75.9 (6.5)	68.8 (5.9)	91.9 (2.2)
XGB	81.9 (6.9)	78.7 (3.5)	79.8 (4.1)	94.6 (2.3)
GRU	65.5 (10.4)	84.4 (2.4)	64.2 (11.6)	94.9 (1.4)
BiGRU	70.4 (10.3)	**89.9 (1.6)**	72.7 (9.3)	**97.6 (1.3)**
BERT	77.3 (6.0)	**93.5 (2.2)**	**82.4 (5.6)**	**98.3 (0.9)**

(a) CSSRS vs. notes

Model	Precision	Recall	F1	AUC
NB	**87.8 (3.2)**	69.2 (1.8)	74.9 (2.1)	93.2 (0.9)
LR	76.4 (2.3)	83.0 (4.9)	79.1 (3.0)	93.2 (1.8)
LSVM	76.4 (1.9)	82.1 (3.0)	78.8 (1.9)	92.3 (1.5)
SVM	74.3 (1.9)	83.9 (3.5)	77.9 (1.8)	93.4 (1.2)
GB	**89.1 (5.7)**	64.2 (3.4)	69.8 (4.2)	91.9 (1.7)
RF	72.3 (1.4)	78.1 (1.9)	74.7 (1.6)	91.3 (0.8)
XGB	76.3 (3.0)	78.4 (2.1)	77.3 (2.5)	93.9 (1.3)
GRU	61.4 (3.6)	83.5 (2.3)	60.8 (6.6)	94.2 (0.9)
BiGRU	64.6 (1.8)	**89.4 (1.0)**	67.5 (2.9)	**97.7 (0.4)**
BERT	79.0 (4.5)	**93.0 (0.7)**	**83.9 (3.8)**	**98.3 (0.4)**

(b) SDD vs. notes

Model	Precision	Recall	F1	AUC
NB	85.3 (1.4)	81.1 (1.0)	81.2 (1.1)	**93.0 (1.9)**
LR	84.5 (4.4)	84.4 (4.5)	84.1 (4.5)	91.4 (2.7)
LSVM	83.6 (3.2)	83.5 (3.4)	83.2 (3.5)	90.8 (2.8)
SVM	84.0 (2.9)	83.7 (2.9)	83.3 (2.9)	90.2 (2.0)
GB	84.6 (2.3)	84.5 (2.2)	84.4 (2.3)	92.6 (2.1)
RF	82.5 (3.4)	82.5 (3.4)	82.3 (3.5)	89.9 (3.3)
XGB	83.6 (3.2)	83.6 (3.4)	83.3 (3.5)	91.3 (3.4)
GRU	80.4 (3.2)	79.4 (2.2)	79.3 (2.4)	89.3 (2.2)
BiGRU	**87.7 (3.2)**	**87.1 (3.2)**	**87.1 (3.2)**	**94.8 (2.1)**
BERT	**86.3 (4.4)**	**86.2 (4.6)**	**86.0 (4.7)**	**93.5 (3.3)**

(c) KSN vs. notes

TABLE IV
MEAN AND STANDARD DEVIATION OF THE 5-FOLD CROSS-VALIDATION RESULTS FOR THE TURKISH-LANGUAGE EXPERIMENTS.

Model	Precision	Recall	F1	AUC
NB	72.1 (7.0)	69.8 (4.9)	70.8 (5.6)	72.6 (7.6)
LR	75.9 (3.1)	74.8 (4.1)	75.3 (3.6)	80.8 (4.8)
LSVM	**79.5 (2.2)**	74.0 (4.2)	**76.2 (3.4)**	83.1 (3.7)
SVM	**79.6 (3.4)**	72.4 (4.7)	75.2 (4.3)	83.6 (3.5)
GB	**79.9 (2.2)**	68.8 (2.5)	72.6 (2.4)	80.7 (4.7)
RF	78.1 (3.1)	68.8 (4.8)	71.8 (3.9)	83.3 (3.1)
XGB	75.2 (4.0)	71.1 (3.3)	72.8 (3.3)	82.7 (4.3)
GRU	59.6 (3.3)	75.8 (8.6)	50.9 (15)	91.9 (2.3)
BiGRU	59.8 (4.7)	75.8 (8.2)	50.7 (13)	**93.1 (1.4)**
BERT	68.3 (2.3)	**87.0 (2.3)**	72.3 (3.1)	**95.0 (1.8)**

(a) TRP vs. notes

Model	Precision	Recall	F1	AUC
NB	80.8 (3.8)	79.6 (1.7)	**80.1 (2.3)**	89.2 (2.4)
LR	83.4 (4.0)	77.9 (3.4)	**80.1 (2.9)**	90.2 (2.9)
LSVM	84.5 (4.8)	77.6 (4.1)	**80.3 (3.6)**	90.5 (3.1)
SVM	84.4 (6.8)	74.1 (5.3)	77.8 (5.7)	88.9 (4.0)
GB	**88.0 (4.2)**	71.8 (3.7)	76.7 (3.2)	89.6 (3.0)
RF	**88.1 (1.9)**	75.0 (3.3)	79.6 (2.4)	91.5 (1.9)
XGB	78.4 (2.7)	77.4 (2.6)	77.6 (0.7)	90.0 (2.0)
GRU	65.8 (1.0)	85.8 (2.1)	67.3 (1.8)	95.2 (0.5)
BiGRU	67.3 (3.6)	86.5 (2.8)	68.9 (6.6)	**96.2 (0.6)**
BERT	75.7 (4.3)	**91.4 (1.4)**	**80.3 (4.2)**	**97.7 (0.6)**

(b) TRM vs. notes

Model	Precision	Recall	F1	AUC
NB	75.7 (9.4)	67.2 (4.6)	69.6 (4.9)	69.3 (6.8)
LR	78.7 (3.1)	73.8 (3.0)	76.0 (2.7)	82.9 (5.3)
LSVM	**81.2** (2.2)	73.3 (3.7)	**76.5** (3.0)	82.5 (8.0)
SVM	76.9 (7.7)	69.4 (6.0)	71.7 (4.8)	78.8 (6.6)
GB	**80.8 (2.3)**	67.6 (4.8)	71.9 (4.7)	78.1 (5.0)
RF	74.4 (3.5)	69.3 (4.2)	71.3 (3.6)	84.6 (5.6)
XGB	78.2 (4.1)	71.6 (3.9)	74.1 (3.5)	84.7 (5.0)
GRU	54.8 (2.1)	70.1 (11)	36.1 (17)	92.3 (2.4)
BiGRU	55.2 (1.6)	76.0 (4.5)	44.1 (6.1)	93.6 (1.8)
BERT	64.7 (2.3)	**89.5 (1.6)**	69.2 (3.1)	**95.9 (0.9)**

(c) TRPM vs. notes

their posts predicted as a suicide note to their total number of posts, and constructed the histogram of these ratios with different colors for the expert-annotated labels. Fig. 6 shows the distribution of these ratios with distinct colors per expert-annotated label. According to the histogram in Fig. 6, a user with the SI label had 90% of posts classified as suicide notes. Similarly, many posts of users with SB and SI labels were classified as a note, where SB bars are taller than the SI bars, which is less frequent for users with the AA label, further validating the previous finding from the confusion matrices.

D. Embedding Analysis

In the final experiment, we visualized the semantic space of document representations using the zero-shot BERT embeddings. In Fig. 7, we mapped the CSSRS, SDD, KSN, and the genuine suicide notes in the same two-dimensional embedding space of the BERT-base model, the best-performing method in English-language experiments. In the figure, there is a tight region underneath the KSN texts that has overlapping SDD and CSSRS data. Meanwhile, there is a region where SDD and CSSRS sparsely overlap, with no KSN data. However, suicide notes appear highly dispersed in the embedding space,

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

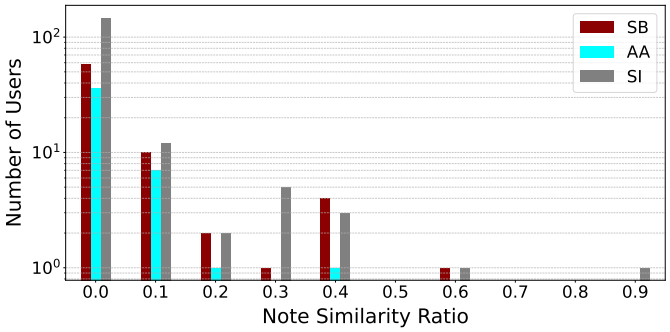


Fig. 6. Histogram of user-level note similarity ratios in the CSSRS dataset.

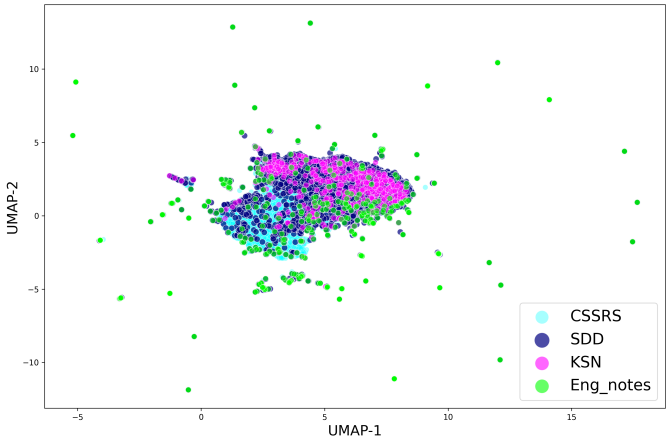


Fig. 7. UMAP projection of English social media posts and suicide notes in the BERT embedding space.

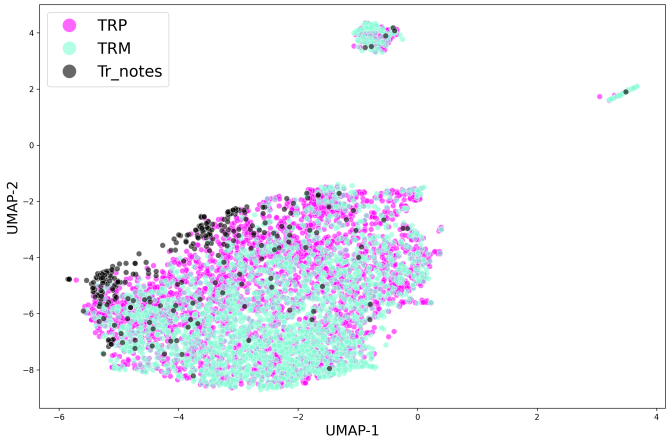


Fig. 8. UMAP projection of Turkish social media posts and suicide notes in the BERTurk embedding space.

with some overlaps with the social media datasets around the edges of the middle area. These observations suggest that KSN semantically overlaps with a large portion of the SDD data and some CSSRS data. But they do not semantically overlap with a large portion of the notes.

Conducting the same experiment using the zero-shot BERTurk embeddings on Turkish data returns Fig. 8, where TRP and TRM posts form an overlapping, indistinct large cluster. Turkish suicide notes, however, appear as a peripheral yet partially integrated group, mostly overlapping with the positive posts at the edge of the large cluster. Also, the zero-shot BERTurk model captured these semantic distinctions and the linguistic cues of suicidal ideation in Turkish.

V. Discussion & Conclusions

The present study had three main objectives: (i) to compare the expression of suicidal ideation in genuine suicide notes and expert-labeled social media datasets using a wide range of methods, (ii) to evaluate the robustness of the assessments across English and Turkish, and (iii) to assess the reliability of heuristically labeled social media data for representing suicidal ideation. To address these objectives, we conducted a series of linguistic, statistical, and computational analyses, ranging from simple machine learning to transformer-based models.

A consistent pattern present across all results is that whether a dataset is manually or heuristically labeled is correlated with the level of the measured alignment with suicide notes. The linguistic analysis showed that expert-labeled data (i.e., CSSRS) has high alignment with suicide notes. In contrast, heuristically labeled or of unknown-source datasets (i.e., SDD and KSN) produced consistent divergence from genuine suicide notes. The fact that CSSRS and SDD originate from the same source, but only differ in labeling strategy, further highlights the effect of annotation strategy.

Additionally, the differences in the first-person pronoun was notable. Studies in the literature have consistently reported elevated usage of this pronoun by English-speaking suicidal individuals compared to the texts generated by non-suicidal individuals [43]–[45]. However, comparing the use of this pronoun in social media posts and genuine suicide notes produced a different outcome: notes had a moderately higher use of this pronoun than the expert-annotated social media corpus, which aligns with the literature. But, the heuristically labeled social media corpus (SDD) and the dataset with an unknown source (KSN) both had higher first-person pronoun use than the notes. This finding highlights a critical limitation of these datasets that lack manual annotations: they might contain more of the self-focused discourse than those with the presence of genuine suicidal ideation. We observe the same characteristic in the religion category of LIWC features, where the CSSRS corpus had the least use of religious words, highlighting that we should be cautious about solely relying on these linguistic features in suicide research.

The n-gram analyses and classification results support the previous findings from the auto-labeled and unknown-sourced corpora by successfully distinguishing them from the notes. However, the classification results contrast with expectations for the expert-labeled dataset: they are correctly separated from the notes despite their previously measured linguistic similarities. Furthermore, this distinguishability is robust across languages as well. This robustness is plausibly due to specific themes present in Turkish and English suicide notes, such as forgiveness, that are absent from social media, reflecting the fact that suicide notes represent a later, more determined stage of suicidal ideation. We also detected demographic differences in the extracted features across these languages:

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

Turkish features contained frequent words referring to children and spouses, whereas English features more often mentioned parents. This difference highlights the importance of selecting or constructing training datasets that represent a wide range of demographics, as models trained on data from a specific age group or demographic may fail to generalize to contexts generated by other groups, with potentially serious consequences in real-world suicide risk detection.

Another notable result is that TRP and TRM share more n-grams than either does with their combination, TRPM, contrary to simple intuitive expectations, while the classification performance of TRPM is lower than that of either dataset alone. After inspecting the confusion matrix of TRPM, we observe that most social media posts are classified correctly, whereas some suicide notes are misclassified as social media posts. Together, these observations suggest that dataset aggregation increased the dominance of social-media-specific linguistic patterns in the models, reducing the separability of note-specific features. Since the models were trained while accounting for class imbalance, this effect is unlikely to be due to label distribution and may instead originate from the mixed feature set. These findings challenge the assumption that larger or more heterogeneous datasets are always more advantageous for classification tasks.

Finally, how the English and Turkish corpora mapped to their respective embedding spaces further clarifies the relationship between suicide notes and social media data across languages. In English, suicide notes are sparsely distributed across a broad region, whereas certain social media datasets cluster with limited overlaps, and KSN has an unusually tight cluster. In Turkish, suicide notes are less dispersed and align more closely with TRP, while TRP and TRM themselves occupy sparser regions with partial overlap. These embedding patterns indicate that although social media data captures some semantic aspects of suicidal ideation in suicide notes, the two data types are not interchangeable.

Overall, our findings highlight both the promise and the limits of using social media data for suicidal ideation detection. While we urgently need large datasets to develop and train models that can support clinicians, reach more populations, and enable early suicide intervention, our results demonstrate that how the data are collected and labeled critically affect what models achieve. Heuristically labeled or unknown-source datasets mostly do not capture the linguistic or semantic characteristics of suicidal ideation expressed in suicide notes, and even expert-annotated social media data only partially overlaps with the ideation in notes, which is consistent across both English and Turkish. These observations do not imply that social media data, or expert-labeled corpora in particular, are unsuitable for suicide risk modeling. Instead, they suggest that suicidal signals in these domains likely correspond to earlier or different stages of ideation than those present in suicide notes. Our findings further caution against assuming that aggregating datasets will necessarily improve performance, as demonstrated by the TRPM corpus.

Our study is subject to several limitations. Our analyses are restricted by two languages and a few available datasets due to the constraint all suicide researchers face: scarcity of ethically sourced, expert-annotated data. Our scripts are available per demand. Future work should extend our study to additional languages and cultural contexts to assess the generality of these findings and to better characterize how suicidal ideation is expressed across linguistic and sociocultural settings. In practice, this may involve prioritizing expert-led or multi-annotator labeling strategies, and restricting the use of automatic labeling as a substitute for human judgment. Encouraged by our findings, we argue that progress in suicide risk detection will depend not only on using larger datasets and developing more powerful models, but on context-aware data collection practices that align annotation strategy, demographic coverage, and the appropriate interpretation of model outputs.

Conflicts of Interest

The authors declare no conflict of interest.

Funding Statement

This study was supported by the Scientific and Technological Research Council of Türkiye (TÜBİTAK) under Grant Number 124E132. The authors thank TÜBİTAK for their support.

References

[1] World Health Organization, "Suicide," Mar. 2025, accessed: 2025-05-15. [Online]. Available: https://www.who.int/news-room/fact-sheets/detail/suicide

[2] S. Ji, S. Pan, X. Li, E. Cambria, G. Long, and Z. Huang, "Suicidal ideation detection: A review of machine learning methods and applications," *IEEE Transactions on Computational Social Systems*, vol. 8, no. 1, pp. 214–226, 2020.

[3] S. S. P. Shukla and M. P. Singh, "Enhancing suicidal ideation detection through advanced feature selection and stacked deep learning models," *Applied Intelligence*, vol. 55, no. 4, p. 303, 2025.

[4] J. Gorai and D. K. Shaw, "A bert-encoded ensembled cnn model for suicide risk identification in social media posts," *Neural Computing and Applications*, vol. 36, no. 18, pp. 10 955–10 970, 2024.

[5] M. Qorich and R. El Ouazzani, "Advanced deep learning and large language models for suicide ideation detection on social media," *Progress in Artificial Intelligence*, vol. 13, no. 2, pp. 135–147, 2024.

[6] J. Pestian, H. Nasrallah, P. Matykiewicz, A. Bennett, and A. Leenaars, "Suicide note classification using natural language processing: A content analysis," *Biomedical informatics insights*, vol. 3, pp. BII–S4706, 2010.

[7] J. P. Pestian, P. Matykiewicz, M. Linn-Gust, B. South, O. Uzuner, J. Wiebe, K. B. Cohen, J. Hurdle, and C. Brew, "Sentiment analysis of suicide notes: A shared task," *Biomedical informatics insights*, vol. 5, pp. BII–S9042, 2012.

[8] A. M. Schoene and N. Dethlefs, "Automatic identification of suicide notes from linguistic and sentiment features," in *Proceedings of the 10th SIGHUM workshop on language technology for cultural heritage, social sciences, and humanities*, 2016, pp. 128–133.

[9] J. Devlin, M. Chang, K. Lee, and K. Toutanova, "BERT: pre-training of deep bidirectional transformers for language understanding," *CoRR*, vol. abs/1810.04805, 2018. [Online]. Available: http://arxiv.org/abs/1810.04805

[10] S. Schweter, "Berturk - bert models for turkish," Apr. 2020. [Online]. Available: https://doi.org/10.5281/zenodo.3770924

[11] G. Coppersmith, R. Leary, P. Crutchley, and A. Fine, "Natural language processing of social media as screening for suicide risk," *Biomedical informatics insights*, vol. 10, p. 1178222618792860, 2018.

[12] M. Gaur, A. Alambo, J. P. Sain, U. Kursuncu, K. Thirunarayan, R. Kavuluru, A. Sheth, R. Welton, and J. Pathak, "Knowledge-aware assessment of severity of suicide risk for early intervention," in *The world wide web conference*, 2019, pp. 514–525.

[13] H.-C. Shing, S. Nair, A. Zirikly, M. Friedenberg, H. Daumé III, and P. Resnik, "Expert, crowdsourced, and machine assessment of suicide risk via online postings," in *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018, pp. 25–36.

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.

This article has been accepted for publication in IEEE Transactions on Affective Computing. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/TAFFC.2026.3664204

[14] B. L. Cook, A. M. Progovac, P. Chen, B. Mullin, S. Hou, and E. Baca-Garcia, "Novel use of natural language processing (nlp) to predict suicidal ideation and psychiatric symptoms in a text-based mental health intervention in madrid," *Comput Math Methods Med*, 2016.

[15] J. Song, T. M. Song, D.-C. Seo, and J. H. Jin, "Data mining of web-based documents on social networking sites that included suicide-related words among korean adolescents," *J Adolesc Health*, vol. 59, no. 6, pp. 668–673, 2016.

[16] H. Woo, Y. Cho, E. Shim, and G. Song, "The sewol ferry disaster and suicide-related public postings," *Int J Environ Res Public Health*, vol. 12, no. 9, pp. 10 974–10 983, 2015.

[17] Z. Li, W. Cheng, J. Zhou, Z. An, and B. Hu, "Deep learning model with multi-feature fusion and label association for suicide detection," *Multimedia systems*, vol. 29, no. 4, pp. 2193–2203, 2023.

[18] V. Desu, N. Komati, S. Lingamaneni, and F. Shaik, "Suicide and depression detection in social media forums," in *Smart Intelligent Computing and Applications, Volume 2: Proceedings of Fifth International Conference on Smart Computing and Informatics (SCI 2021)*. Springer, 2022, pp. 263–270.

[19] K. Nikhileswar, D. Vishal, L. Sphoorthi, and S. Fathimabi, "Suicide ideation detection in social media forums," in *2021 2nd International Conference on Smart Electronics and Communication (ICOSEC)*. IEEE, 2021, pp. 1741–1747.

[20] A. M. Schoene, L. Bojanić, M.-Q. Nghiem, I. M. Hunt, and S. Ananiadou, "Classifying suicide-related content and emotions on twitter using graph convolutional neural networks," *IEEE Transactions on Affective Computing*, vol. 14, no. 3, pp. 1791–1802, 2022.

[21] S. Ji, C. P. Yu, S.-f. Fung, S. Pan, and G. Long, "Supervised learning for suicidal ideation detection in online user content," *Complexity*, vol. 2018, no. 1, p. 6157249, 2018.

[22] S. Mankarious, A. Zirikly, D. Wiechmann, E. Kerz, E. Kempa, and Y. Qiao, "Mindset: Advancing mental health benchmarking through large-scale social media data," *arXiv preprint arXiv:2511.20672*, 2025.

[23] T. Zhang, A. M. Schoene, and S. Ananiadou, "Automatic identification of suicide notes with a transformer-based deep learning model," *Internet interventions*, vol. 25, p. 100422, 2021.

[24] J. Pestian, D. Santel, M. Sorter, U. Bayram, B. Connolly, T. Glauser, M. DelBello, S. Tamang, and K. Cohen, "A machine learning approach to identifying changes in suicidal language," *Suicide and Life-Threatening Behavior*, vol. 50, no. 5, pp. 939–947, 2020.

[25] E. Lin, J. Sun, H. Chen, and M. H. Mahoor, "Data quality matters: Suicide intention detection on social media posts using roberta-cnn," in *2024 46th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*. IEEE, 2024, pp. 1–5.

[26] A. M. Schoene, A. P. Turner, G. De Mel, and N. Dethlefs, "Hierarchical multiscale recurrent neural networks for detecting suicide notes," *IEEE Transactions on Affective Computing*, vol. 14, no. 1, pp. 153–164, 2021.

[27] U. Naseem, M. Khushi, J. Kim, and A. G. Dunn, "Hybrid text representation for explainable suicide risk identification on social media," *IEEE transactions on computational social systems*, 2022.

[28] R. S. Soun, A. T. Neerkaje, R. Sawhney, N. Aletras, and P. Nakov, "Rise: Robust early-exiting internal classifiers for suicide risk evaluation," in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 2024, pp. 14 134–14 145.

[29] Ö. Ezerceli and R. Dehkharghani, "Mental disorder and suicidal ideation detection from social media using deep neural networks," *Journal of Computational Social Science*, vol. 7, no. 3, pp. 2277–2307, 2024.

[30] S. R. Lobo and S. Kamath, "Leveraging text analysis and machine learning models for suicide ideation detection in social media posts," *SN Computer Science*, vol. 6, no. 6, p. 703, 2025.

[31] S. Ghosh, A. Ekbal, and P. Bhattacharyya, "Cease, a corpus of emotion annotated suicide notes in english," in *Proceedings of the twelfth language resources and evaluation conference*, 2020, pp. 1618–1626.

[32] ——, "Deep cascaded multitask framework for detection of temporal orientation, sentiment and emotion from suicide notes," *Scientific reports*, vol. 12, no. 1, p. 4457, 2022.

[33] S. Ji, X. Li, Z. Huang, and E. Cambria, "Suicidal ideation and mental disorder detection with attentive relation networks," *Neural Computing and Applications*, 2021.

[34] Y. R. Tausczik and J. W. Pennebaker, "The psychological meaning of words: Liwc and computerized text analysis methods," *Journal of language and social psychology*, vol. 29, no. 1, pp. 24–54, 2010.

[35] X. S. Liu, *Statistical power analysis for the social and behavioral sciences: Basic and advanced techniques*. Routledge, 2013.

[36] J. W. Pennebaker, C. K. Chung, M. Ireland, A. Gonzales, and R. J. Booth, "The development and psychometric properties of liwc2007," LIWC.net, Austin, TX, Technical Report, 2007.

[37] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg *et al.*, "Scikit-learn: Machine learning in python journal of machine learning research," *Journal of machine learning research*, vol. 12, pp. 2825–2830, 2011.

[38] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga *et al.*, "Pytorch: An imperative style, high-performance deep learning library," in *Proceedings of the 33rd Conference on Neural Information Processing Systems (NeurIPS 2019)*. Vancouver, Canada: NeurIPS, 2019, pp. 8024–8035.

[39] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, "Optuna: A next-generation hyperparameter optimization framework," in *Proceedings of the 25th ACM SIGKDD international conference on knowledge discovery & data mining*, 2019, pp. 2623–2631.

[40] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *Proceedings of the 22nd acm sigkdd international conference on knowledge discovery and data mining*, 2016, pp. 785–794.

[41] S. Sari and U. Bayram, "Vaccine hesitancy in Türkiye: A natural language processing study on social media," *Turkish Journal of Electrical Engineering and Computer Sciences*, vol. 33, no. 3, pp. 392–409, 2025.

[42] L. McInnes, J. Healy, and J. Melville, "Umap: Uniform manifold approximation and projection for dimension reduction," *arXiv preprint arXiv:1802.03426*, 2018.

[43] V. Venek, S. Scherer, L.-P. Morency, J. Pestian *et al.*, "Adolescent suicidal risk assessment in clinician-patient interaction," *IEEE Transactions on Affective Computing*, vol. 8, no. 2, pp. 204–215, 2017.

[44] C. Lao, J. Lane, and H. Suominen, "Analyzing suicide risk from linguistic features in social media: evaluation study," *JMIR formative research*, vol. 6, no. 8, p. e35563, 2022.

[45] S. Homan, M. Gabi, N. Klee, S. Bachmann, A.-M. Moser, S. Michel, A.-M. Bertram, A. Maatz, G. Seiler, E. Stark *et al.*, "Linguistic features of suicidal thoughts and behaviors: A systematic review," *Clinical psychology review*, vol. 95, p. 102161, 2022.

[46] A. Lex, N. Gehlenborg, H. Strobelt, R. Vuillemot, and H. Pfister, "Upset: visualization of intersecting sets," *IEEE transactions on visualization and computer graphics*, vol. 20, no. 12, pp. 1983–1992, 2014.

[47] J. T. Hancock, T. M. Khoshgoftaar, and J. M. Johnson, "Evaluating classifier performance with highly imbalanced big data," *Journal of Big Data*, vol. 10, no. 1, p. 42, 2023.



Merve Kavut is an M.Sc. student in the Dept. of Electrical and Electronics Engineering at Çanakkale Onsekiz Mart University (ÇOMÜ), Türkiye. She received her B.Sc. in Computer Engineering from ÇOMÜ in 2024. She is funded by the TÜBİTAK project under Grant Number 124E132. Her research interests include NLP, and complex networks.



Amina Dzafic is an M.Sc. student in the Dept. of Electrical and Electronics Engineering at ÇOMÜ, Türkiye. She received her B.Sc. in Computer Engineering from ÇOMÜ in 2023. She is funded by the TÜBİTAK project under Grant Number 124E132. Her research interests include Data Science, NLP, and machine learning.



Ulya Bayram is an Assistant Professor in the Dept. of Electrical and Electronics Engineering at ÇOMÜ, Türkiye. She received her Ph.D. degree in Computer Science from the University of Cincinnati, Cincinnati, OH, USA, in 2019, where she conducted research at Cincinnati Children's Hospital Medical Center focusing on suicide-related NLP. She is the Principal Investigator of several TÜBİTAK-funded research projects.

Authorized licensed use limited to: Canakkale 18 Mart Universitesi. Downloaded on April 11,2026 at 12:47:46 UTC from IEEE Xplore. Restrictions apply.
© 2026 IEEE. All rights reserved, including rights for text and data mining and training of artificial intelligence and similar technologies. Personal use is permitted, but republication/redistribution requires IEEE permission. See https://www.ieee.org/publications/rights/index.html for more information.